Dec 2018 DataOne Webinar

A Shared Staffing Model for Research Data Curation

Lisa Johnston
University of Minnesota

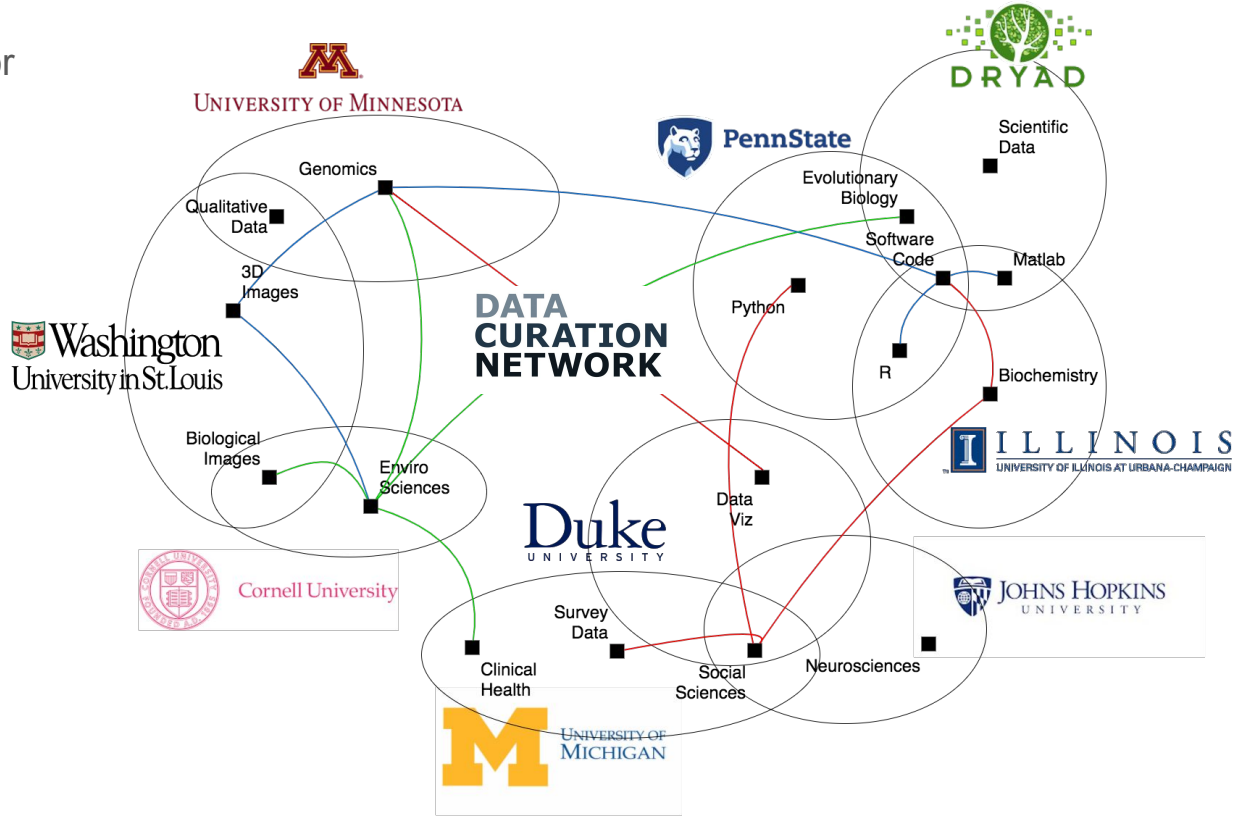
Debra Fagan
Dryad Data Repository

2018-12-11

DataCurationNetwork.org

DATA CURATION NETWORK

- Collaborative staffing model for curating research data across academic and general data repositories
- Funded by Alfred P. Sloan Foundation
- Implementation phase (2018-2021)
  - 8 member institutions
  - 32 individuals
- Goal is to expand to all users in 2020



Well curated data are more valuable.

- Easier for fellow scholars and future collaborators to understand
- More likely to be trusted and reused
- The research findings they represent are more likely to be validated and replicated
- More likely to be properly cited

Due to the heterogeneous and multidisciplinary nature of research data generated by the researchers we support, data curation can be a challenge.

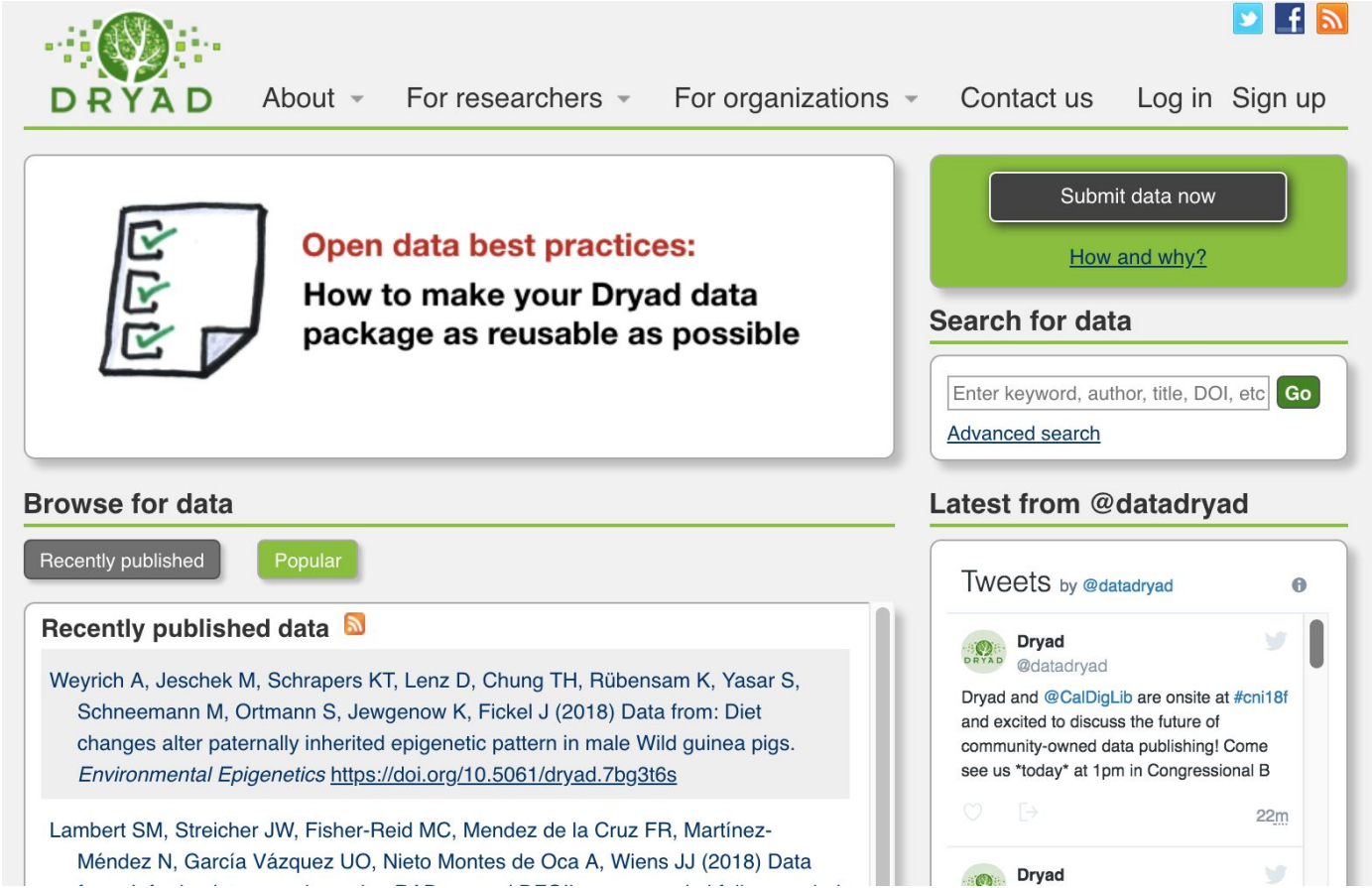
The Data Curation Network (DCN) addresses this challenge by collaboratively sharing data curation staff across a network of partner institutions and data repositories.

DRYAD

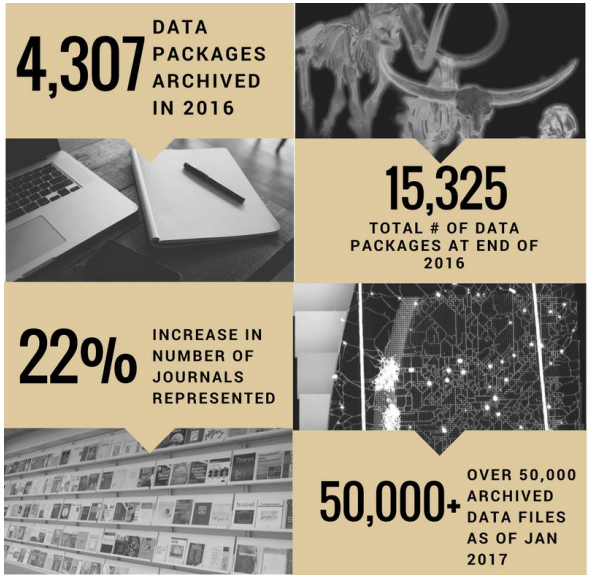
Nonprofit, curated general-purpose repository for data underlying scholarly publications

Goal - make data discoverable, freely reusable and citable

Evolved from ecology and evolutionary bio data -> diverse data in a variety of disciplines



Dryad - Curation Challenges



Numerous challenges:

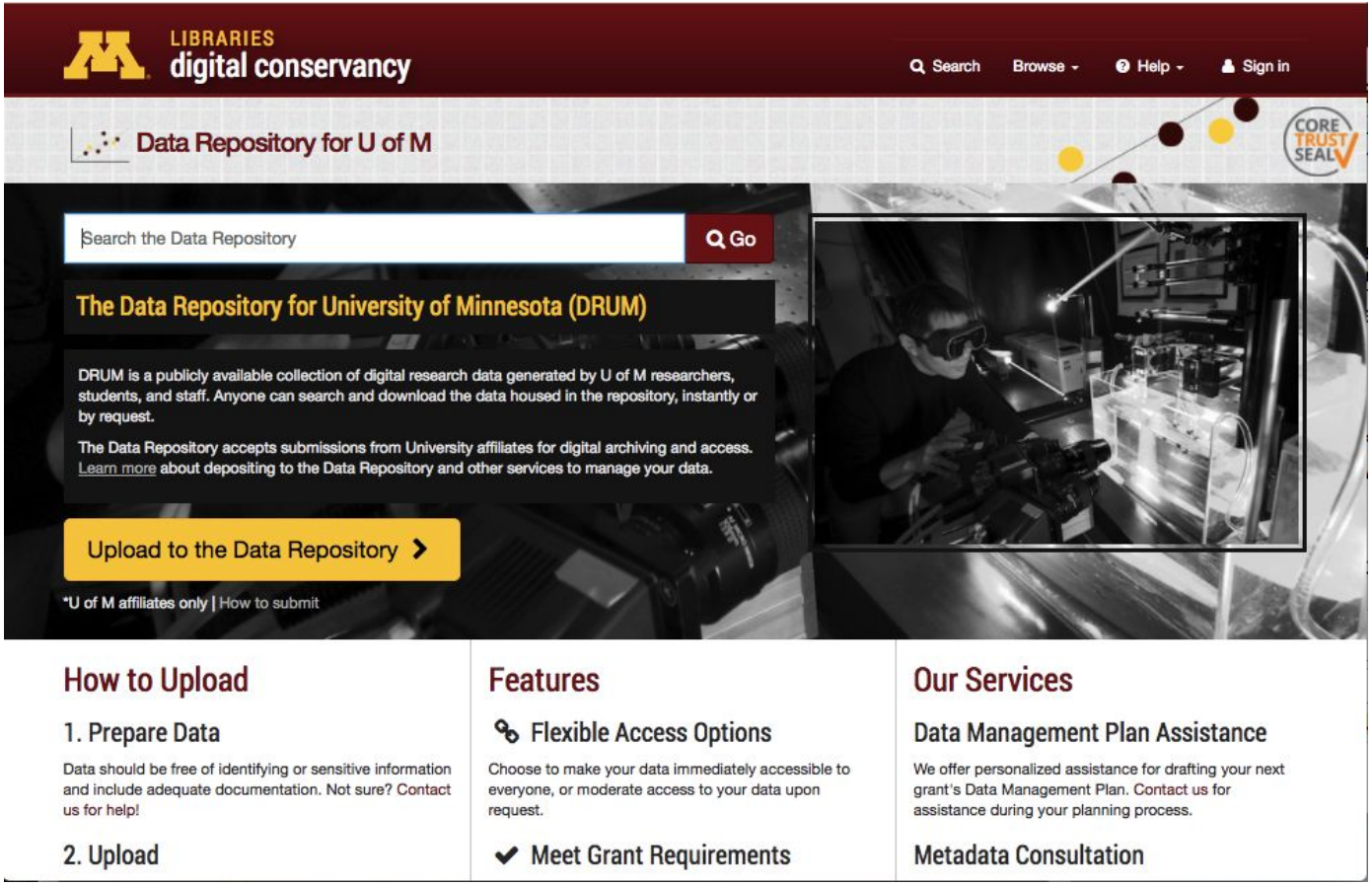
- Licensing / Copyright
- Human subject information
- Endangered species
- Discipline specific knowledge
- Skills and resources needed to curate a wide variety of data

DRUM

Mediated deposit with post-ingest curation

7 data curation staff
- Scientific (2)
- Spatial/GIS
- Public Health
- Digital Hum
- Social Sciences
- Coordinator

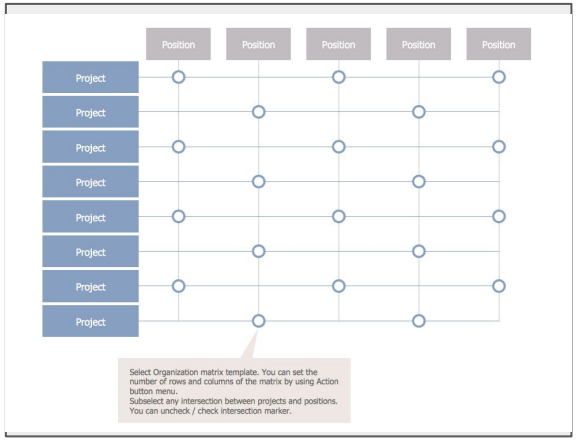
Part of a larger RDS service run by the libraries

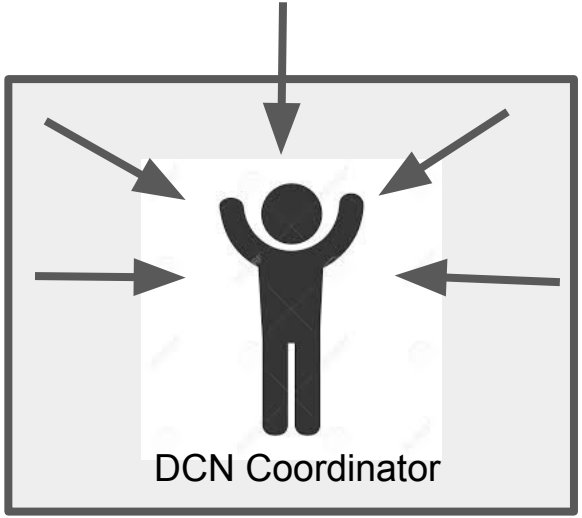


Project → Service

1. *How do we get the individual datasets to the right data curation expert?*
2. *How do we account for differences across institutions with differing policies/infrastructure?*
3. *How do we create consistency across different curators?*
4. *How do we communicate effectively?*
5. *How do we sustain and grow?*

How do we get the individual datasets to the right data curation expert?





How do we account for differences across institutions with differing policies/infrastructure?

1

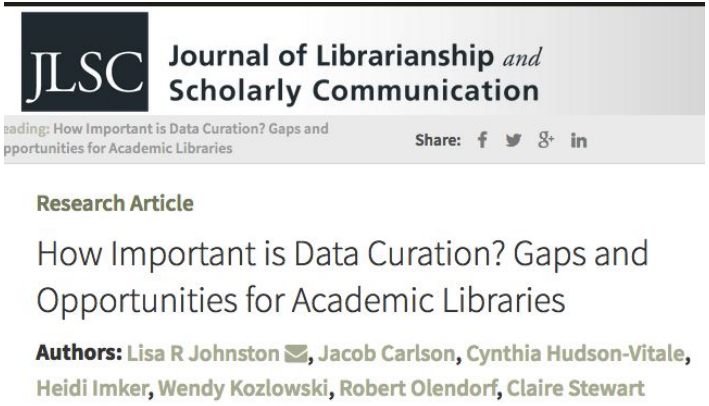
Understand your partners! Baseline assessment and researcher focus groups at each institution

Journal of eScience Librarianship ISSN 2161-3974

Volume 6 | Issue 1 Article 3

2017-02-28

Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions' Data Repository and Curation Services



JLSC Journal of Librarianship and Scholarly Communication

eading: How Important is Data Curation? Gaps and pportunities for Academic Libraries Share:

Research Article

How Important is Data Curation? Gaps and Opportunities for Academic Libraries

Authors: Lisa R Johnston, Jacob Carlson, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert Olendorf, Claire Stewart

Unique Perspectives and Infrastructure

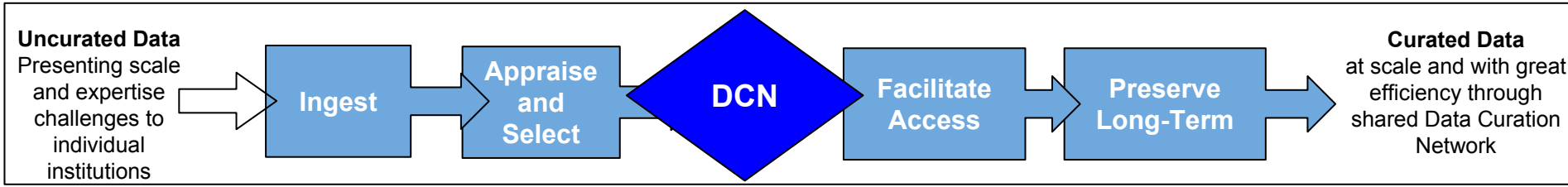
Institution	Data Repository	Launch Date	Dataset Holdings (as of April 2018)	Technology Platform
Cornell University	eCommons	Fall 2002	120 datasets	DSpace 6.2
Dryad Digital Repository	Dryad Digital Repository	2009	21,293 data packages; 67,907 individual files	DSpace moving to CDL's Merritt
Duke University	Duke Digital Repository (DDR)	Jan 2017	34 data deposits total: 8,120 files (most average 50-100)	Fedora 3.8, and Hyrax
Johns Hopkins University	JHU Data Archive	2011	61 datasets or dataverses; 448 files	Dataverse
Penn State University	ScholarSphere	Fall 2012	966 files	Hydra/Fedora (soon to be Sufia 7)
University of Illinois	Illinois Data Bank	May 2016	77 datasets	Ruby on Rails interface with our preservation repository (Medusa)
University of Michigan	Deep Blue Data	Sep 2016	114 datasets	Samvera/Fedora Hyrax
University of Minnesota	Data Repository for the U of M (DRUM)	Mar 2015	199 in DRUM, ~600 in IR	DSpace 5.5

How do we account for differences across institutions with differing policies/infrastructure?

Understand your partners! Baseline assessment and researcher focus groups at each institution

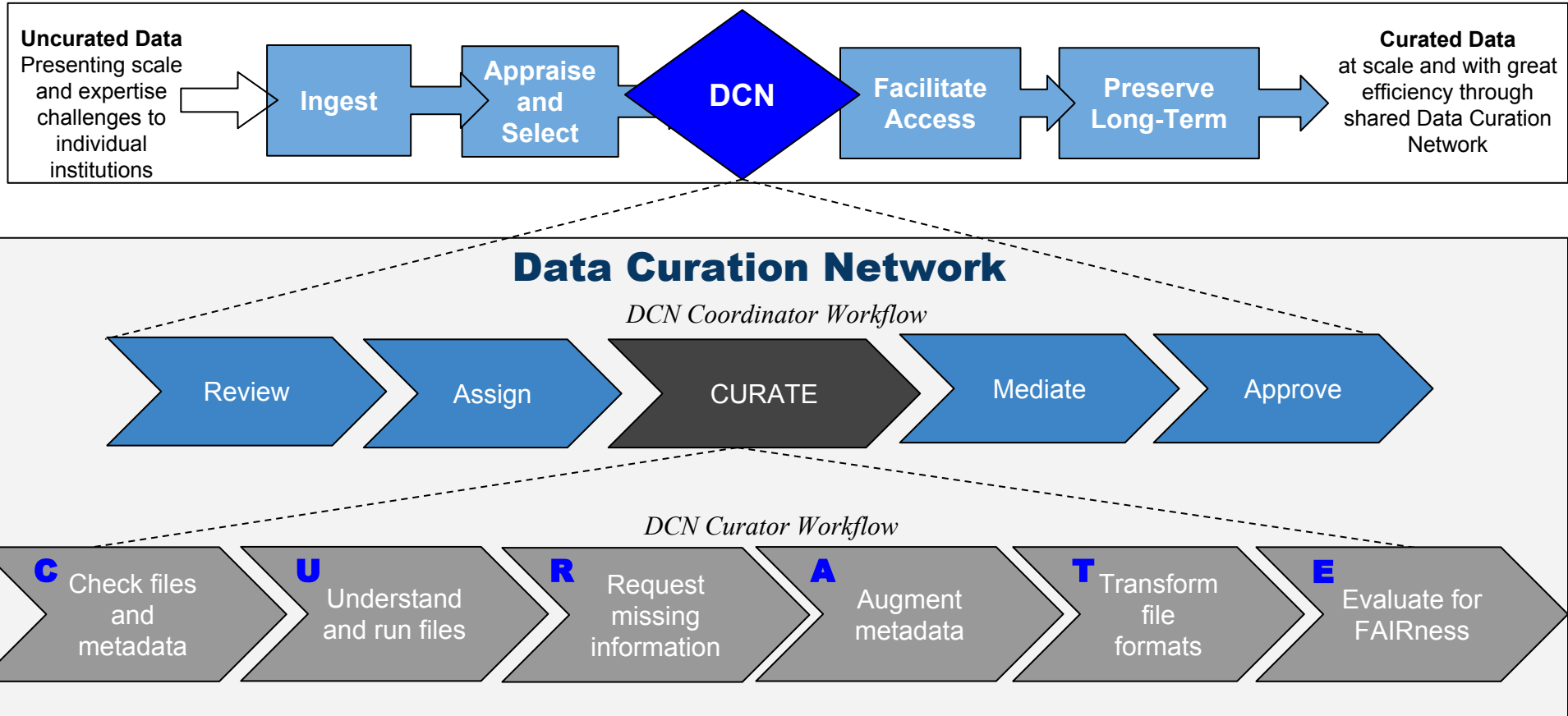
DCN workflow with built-in flexibility for local interpretation

DCN Workflow

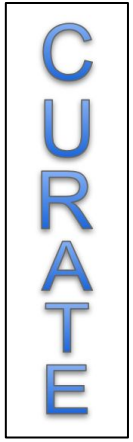


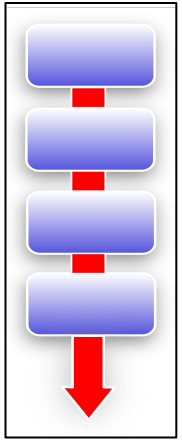
- Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)
- Local institution maintain full responsibility for all technical functionality (eg. storage) and authority for local decision-making (what to ingest, how long to retain, etc.)
- Seamlessly integrates into all repository systems (Samvera, Fedora, DSpace, etc.)

DCN Workflow



How do we account for differences across individual curators?





CURATE Steps in DCN Workflow

DCN Curators will take **CURATE** steps for each data set, that includes:

- **C** **Check** data files and read documentation
- **U** **Understand** the data (try to), if not…
- **R** **Request** missing information or changes
- **A** **Augment** the submission with metadata for findability
- **T** **Transform** file formats for reuse and long-term preservation
- **E** **Evaluate** and rate the overall submission for FAIRness.

DCN CURATE Checklists

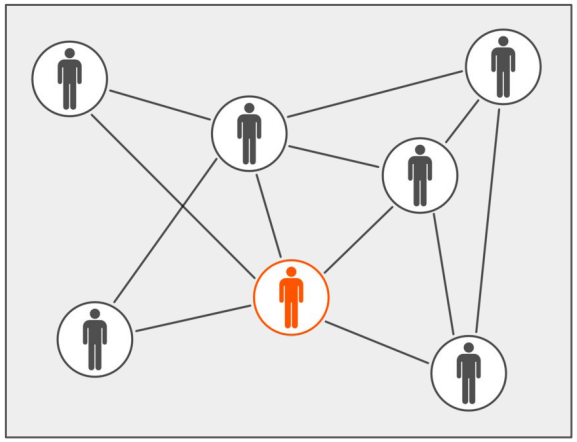
Table A1. Draft checklist of DCN CURATE steps and FAIRness scorecard

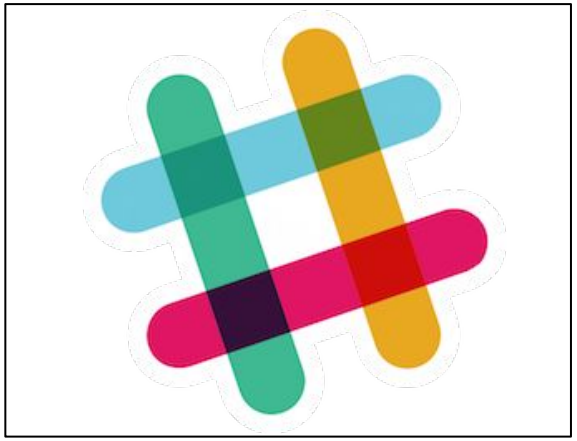
CURATE Actions	Curation Check
Check data files and read documentation ▪ Review the content of the data files (e.g., open and run the files or code). ▪ Verify all metadata provided by the author and review the available documentation.	☐ Files open as expected ☐ Issues _____ ☐ Code runs as expected ☐ Produces minor ☐ Does not run an many errors ☐ Metadata quality is rich, complete ☐ Metadata has iss ☐ Documentation Type (*ci* Readme / Codebook / D Other: _____ ☐ Missing/None ☐ Needs work
Understand the data (or try to) ▪ Check for quality assurance and usability issues such as missing	*Varies based on file formats and s* *example….*
Evaluate and rate the overall data record for FAIRness.[2] ▪ Score the dataset and recommend ways to increase the FAIRness of the data and become "DCN approved."	Findable - ☐ Metadata exceeds author/ title/ date, ☐ Unique PID (DOI, Handle, PURL, etc.). ☐ Discoverable via web search engines like Google. Accessible - ☐ Retrievable via a standard protocol (e.g., HTTP). ☐ Free, open (e.g., download link). Interoperable - ☐ Metadata formatted in a standard schema (e.g., Dublin Core). ☐ Metadata provided in machine-readable format (OAI feed). Reusable - ☐ Data include sufficient metadata about the data characteristics to reuse without the direct assistance of the author. ☐ Clear indicators of who created, owns, and stewards the data. ☐ Data are released with clear data usage terms (e.g., a CC License).

[1] Format Recommendations, http://guides.library.cornell.edu/ecommons/formats

[2] Rubric evaluating the FAIR principles are based on the scoring matrix by Dunning, de Smaele, & Böhmer (2017).

How do we communicate effectively?





How do we sustain and grow?

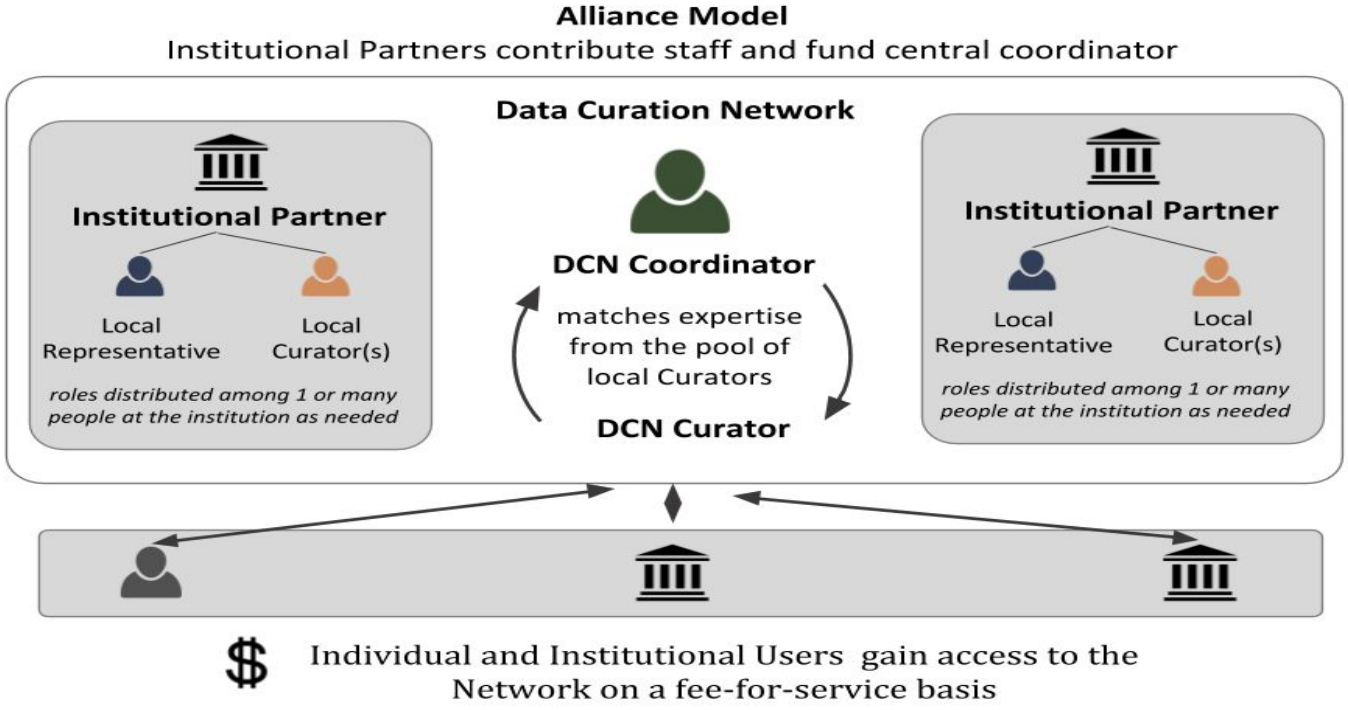
1 *Train new curators (IMLS grant!)*

DATA CURATION WORKSHOP

INSTITUTE of Museum and Library SERVICES

How do we sustain and grow?

2



DCN Research Questions

Is a networked approach to curating research data more efficient?

- Number of datasets
- Frequency (high-volume time periods, etc.)
- Variety (data file formats; range of disciplines)
- Efficiency (time, costs)
- Curator incentives?

Are curated data are more valuable?

- Survey researcher satisfaction
- Track reuse indicators (download counts, citations, alt-metrics)
- Implement a DCN registry
- Apply badges and metadata to signal that data sets curated by the DCN are FAIR.

Thanks!

https://DataCurationNetwork.org

Twitter: #DataCurationNetwork